Comprehensive Centralized-Data Warehouse for Managing Malaria Cases

Nova Eka Diana
YARSI E-Health Research Center
Faculty of Information Technology
YARSI University
Jakarta, Indonesia

Aan Kardiana
YARSI E-Health Research Center
Faculty of Information Technology
YARSI University
Jakarta, Indonesia

***Abstract*—Tanah Bumbu is one of the most endemic areas in Indonesia for patients diagnosed with malaria diseases. Currently, available malaria case data were stored in disparate sources. Hence, it is difficult for the public health department to quickly and easily gather the useful information for determining strategic actions in tackling these cases. The purpose of this research is to build a data warehouse that integrates all malaria cases from disparate sources. This malaria data warehouse is a centralized architecture of galaxy or constellation scheme that consists of three fact tables and 13 dimension tables. SQL Server Integration Services (SSIS) is utilized to build ETL packages that load data from various sources to stages, dimensions, and fact tables in malaria data warehouse. Finally, a timely report can be generated by extracting the salient information located in malaria data warehouse.**

Keywords—malaria case; centralized data warehouse; galaxy scheme; ETL; timely report

I. Introduction

Malaria is one of infectious disease spread by a mosquito of the genus Anopheles. This animal carries out a plasmodium parasite and spread it into human blood circulation through a bite. Every year, there are about 300-500 million people infected by malaria and 1 million people died because of this disease. According to Global Health Observatory (GHO) data, in 2012, there were an estimated about 207 million cases of malaria worldwide and most of those (about 80%) occurred in sub-Saharan Africa. In this report, Indonesia, one of the endemic countries, was mentioned to have about 343,527 cases of malaria with 45 people reported to be dead in 2013 [1]. The number of malaria incidences in Indonesia has decreased by 2.9% in 2007 to 1.9% in 2013. However, contrary to these conditions, the number of cases in West Papua has increased quite sharply in 2013. This area is located in the east part of Indonesia with the most prevalence number above the average [2]. One possible reason for these phenomena is the fairness of medical supply distribution. The scattering locations of malaria incidence may affect a different responsiveness over the cases treatments. Lack of centralized malaria data all over Indonesia could be the reason as well.

Data warehouse has been widely used to manage a significant volume of data that spread in scattered locations. Data warehousing system can be considered as a collection of methods, techniques, and tools to assist managerial users, e.g. senior managers and directors, to administer their jobs. Data warehouse could provide some salient information that helps these users to conduct data analysis in decision-making processes [3]. In recent years, healthcare industry and organization have started adopting a predictive analytic approach for a variety of purposes. To support this idea, they must develop an infrastructure that able to generate timely reports and intervention strategies for health care problems. Healthcare organization needs to build an advanced data warehouse that integrates all available information in a real time manner so that can accommodate those capabilities. A proper deployment of successful data warehouse in managing diseases information will benefit both the organizations and the patients [4-5].

In this article, data warehouse building is proposed to manage and integrate a scattered data of Malaria cases in Tanah Bumbu, one of the endemic areas in Indonesia. By using this proposed data warehouse, the public health department can extract and generate information that useful for decision-making processes. They also can generate and visualize timely reports to present information in an interactive way that easier for executives in the public health department to understand.

The rest of this article is organized as following. Section 2 describes literature review about methodologies used to develop a data warehouse. Next, Section 3 explains the centralized architecture and Single Dimensional Data Store (DDS) scheme used to build malaria data warehouse. In Section 4, data collection and analysis are conducted to determine the stages, dimensions, and fact tables for malaria data warehouse. After that, ETL packages for translating various data sources into a single data warehouse are depicted in Section 5. Finally, the conclusion and future works are defined in Section 6.

II. Data Warehouse Methodology

Data warehouse can be considered as a central repository of information that integrates various data from one or more disparate sources. Data warehouse is usually described as a collection of subject-oriented, integrated, non-volatile, and time-varying data to support decision-making processes [6]. Data warehouse gives a multidimensional view of a big amount of historical data from operational data sources to provide useful information for decision maker in improving their organization business process [7].

This research is funded by Unggulan Perguruan Tinggi (UPT) grant from Directorate General of Higher Education (DIKTI) of Indonesia Ministry of Education and Culture.

WH Inmon made an observation on classical system development life cycle (SDLC) which assumes that requirements are identified at the beginning of design process of Decision Support System (DSS). However, in the real practices, requirements are usually the last one discovered in development process [6]. Inmon pointed out that data warehouse design was a data-driven approach of which analyst frequently understood requirements and data that available for them, only after they encountered opportunities to perform various kinds of data analysis.

Data warehouse development process is a cycle rather than a serialized time and it repeats every 12 to 18 months [8]. This cycle consists of five major steps as illustrated in Fig. 1.

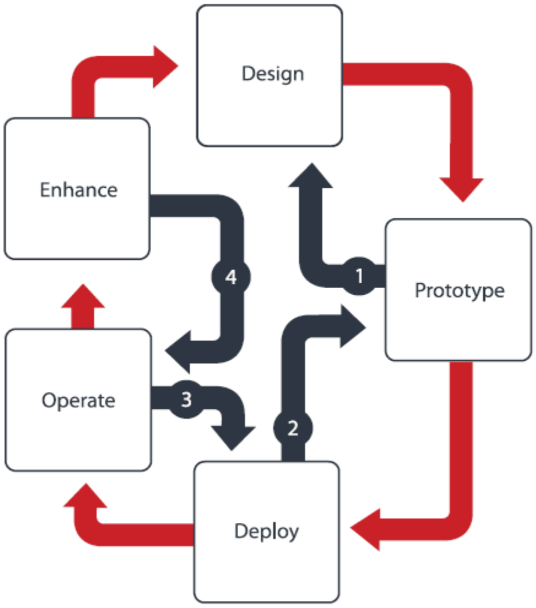


Fig. 1. Data Warehouse Lifecycle Model [8]

Those five major steps are:

- Design. In this phase, the developers create a robust dimensional data models based on available data and analyst requirements.
- Prototype. The main objective of this stage is to constrain and reframe end-user requirements by giving a group of decision-makers and leading practitioners what they need.
- Deploy. There are at least two separate deployment processes should be conducted in the development process: deployment of a prototype to production-test environment and performance-tested production to an actual production environment.
- Operate. Here, the developer conducted a day-to-day maintenance of the data warehouse.
- Enhance. This step includes modification of physical components, operations and management processes in response to business requirement changes.

III. Data Warehouse Infrastructure

A. Data Warehouse (DW) Architecture

Five dominant architectures that are usually adopted to build DW infrastructure: independent data mart, hub-and-spoke, bus, centralized and federated architecture [9-11]. Among these architectures; bus, hub-and-spoke, and centralized, are equally successful for their intended purposes. There is no single dominant architecture in terms of information and system quality, individual and organizational impacts. There is no clear winner among these architecture designs. The differences among them lie on cost, adaptability, scalability, and efficiency of an organizational business process. Therefore, IT managers should consider those factors in deciding the right architecture for building their data warehouse infrastructure [12-13].

In this research, the centralized architecture of malaria data warehouse was proposed as illustrated in Fig. 2. Many applications can directly access this data warehouse to extract salient information for advanced purposes. The whole processes executed in centralized architecture are depicted in Fig. 3. Data warehouse system collects raw data from various data sources such as a database (DBMS), spreadsheet and CSV files. All of these data are then transformed into a uniform format and stored in one place called as Data Stage. After that, these data will be distributed to various components of data warehouse storage, e.g. dimensional and fact tables. When disparate data sources have resided in malaria data warehouse, then different applications can connect to provide any services related to decision support making.

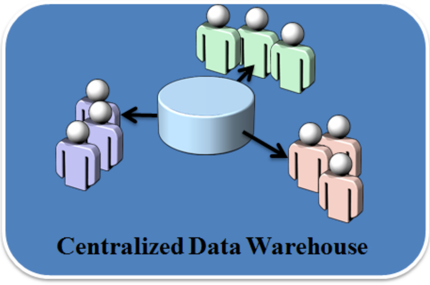


Fig. 2. Centralized Data Warehouse

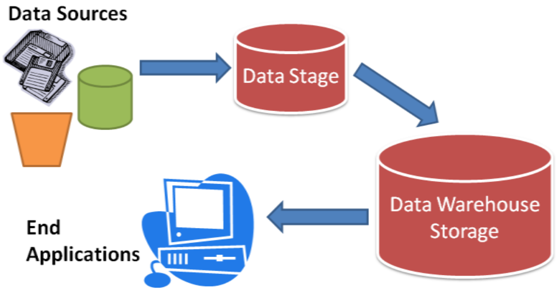


Fig. 3. Processes in centralized architecture

B. Data Flow Architecture

Here, Single Dimensional Data Store (Single DDS) was employed to build malaria data warehouse as shown in Fig. 4.

In this data flow, ETL packages gathered data from various source systems and move it to Stage data store. Next, DDS ETL packages and Data Quality (DQ) would extract data from Stage data store and distribute it to a correct DDS in the data warehouse. Data resided in DDS were then accessed to provide useful information for various applications. Throughout these processes, control-audit administered all ETL packages based on the structure of the data and the description of the processes saved in a metadata.

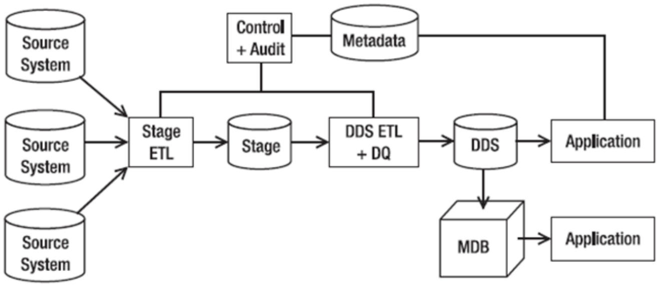


Fig. 4. Single DDS Data Flow Architecture [14]

IV. Data Collection and Analysis

Data sources used in this research were a collection of patients' data diagnosed with malaria in Tanah Bumbu Regency, South Kalimantan, one of the most endemic areas in Indonesia. These data sources consisted of numerous spreadsheet files that recorded malaria cases from 2012 to 2014. Each file contained information about the status of patients' medication for one month period. Therefore, there would be twelve spreadsheet files that recorded malaria cases in a year. Moreover, the medication process for a patient diagnosed with malaria might need more than one month. Hence, it was also possible for one patient' data to be recorded in more than one spreadsheet file.

Here, malaria data were analyzed to identify entities, attributes, and relationship among the entities. Based on the data behavior, those entities were categorized into dimension and fact tables. In this process, numerous spreadsheet files were transformed into a uniform format and then located in a staging table. There were two stage tables created in this data warehouse, Stage of Malaria Data and Stage of Area Information. Stage of Malaria Data recorded patients' data that were diagnosed with malaria. Stage of Area Information stored data about areas and location of the health services that managed Malaria cases.

Fig. 5 illustrates the relationship among stages, dimensions, and fact tables of the malaria data warehouse. Here, more than one fact table and one-dimensional table are shared by many fact tables. Hence, it is called as Galaxy or Fact-Constellation Scheme. In this malaria data warehouse, there are three fact tables: Fact of Malaria Data, Fact of Area Statistic, and Fact of Logistic. Fact of Malaria Data records rows of data related to patient information, type of malaria disease, laboratory results, and type of services given to patients. Fact of Area Statistic contains data that can be used to analyze the starting point and the spread of malaria disease. Fact of Logistic records medicines and supplies that are distributed to each health services that responsible for handling patients diagnosed with malaria. This table can also be used to analyze and predict the utilization of some medicines and supplies for each health services location in each period. Three fact tables, Fact of Malaria Data, Fact of Area Statistic, and Fact of Logistic tables, shared Geographic Dimensional table to track the location of patient and health services area.

V. Results and Discussion

After designing the scheme structure of Malaria data warehouse, data sources from numerous spreadsheet files were converted and formatted into a uniform format. Multiple ETL packages were created using SQL Server Integration Services (SSIS) to populate data to malaria data warehouse. These ETL packages were classified into three categories: Stage ETL, Dimension ETL, and Fact ETL packages.

A. Stage ETL Packages

Stage ETL packages were created to populate data sources into stage tables in the data warehouse. Stage tables are merely just a regular tables, but they have a role in storing rows of data from various data sources. Hence, they can be accessed quickly. Two ETL packages were created for populating data into StageDataMalaria and StageInfoWilayah from spreadsheet data sources. Fig. 6 and Fig. 7 depicted data flow process of Stage of Malaria Data and Stage of Area Information ETL. Stage of Malaria Data ETL consisted of three components: data source, data flow transformation, and data destination. Data flow transformation in this ETL was a derived column type to generate a new value by applying some expressions. ISNULL expression was utilized here to convert number into string values as shown in Table I.

TABLE I. Data Conversion Example

Value	Generated Value
NULL	""
1	PUSTU
2	Poskesdes
3	Polindes/Bidan Desa
4	Klinik/Praktek Swasta
5	Kader/Posmaldes

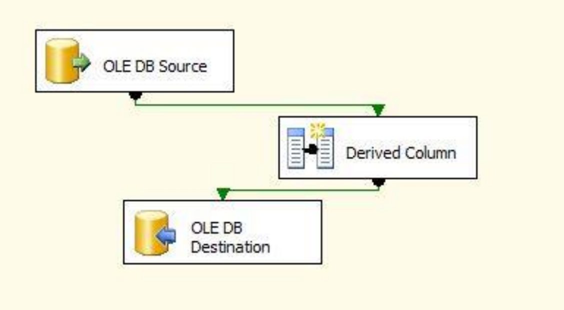


Fig. 5. Stage of Malaria Data ETL

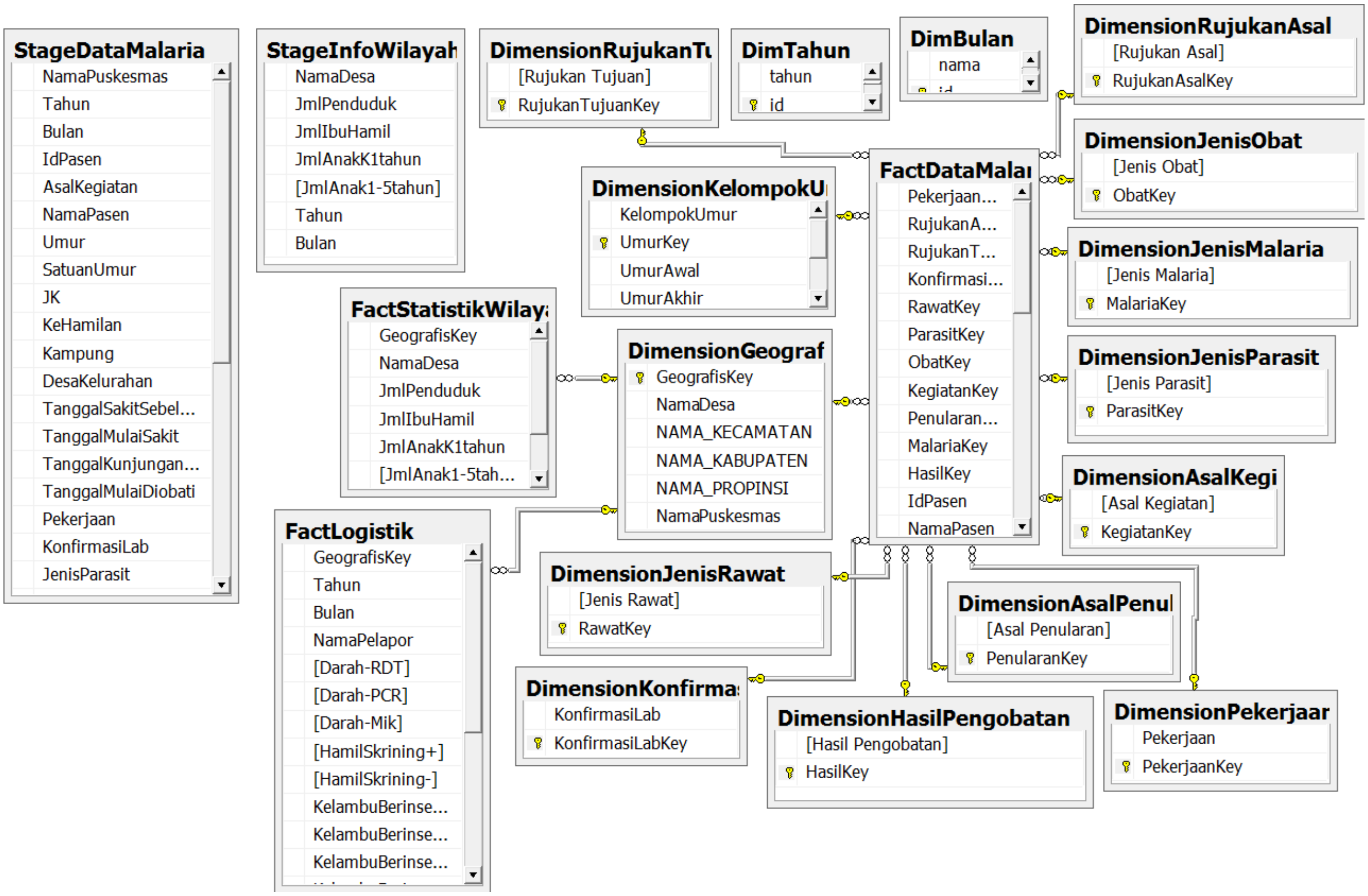


Fig. 6. Galaxy Scheme of Malaria Data Warehouse

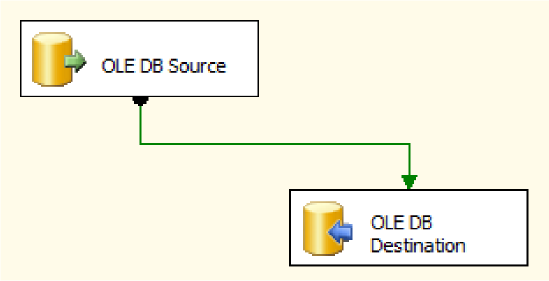


Fig. 7. Stage of Area Information ETL

B. Dimensional ETL Packages

All dimensional ETL packages transform the master data sources into the dimensional tables. Data from dimensional and stage tables will be utilized by ETL packages to fill in the fact tables in the data warehouse. In this process, there are 13 ETL packages created to fill in dimensional tables in Malaria data warehouse as depicted in Table II.

Fig. 8 illustrates the process of executing ETL to populate data into Dimension of Medication Result. In this process, ETL connects to master data sources, extract the medication result and then populate them into Dimension of Medication Result. Fig. 9 shows data in Dimension of Medication Result after successfully execute the ETL. ETL package is successfully executed when the color of each component is green and displays the number of data transferred from OLE DB Source to OLE DB Destination. In this case, four rows of data are successfully populated into Dimension of Medication Result.

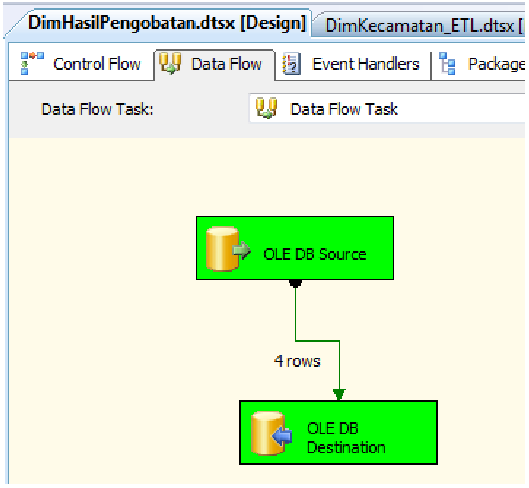
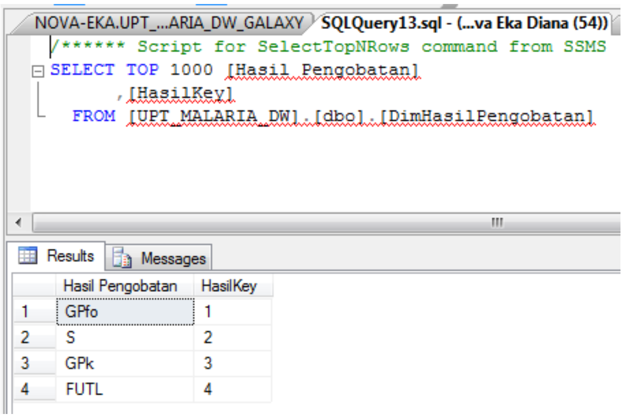


Fig. 8. Dimension of Medication Result ETL execution

TABLE II. DIMENSIONAL ETL PACKAGES

No	ETL Name
1	Dimension of Destination Reference ETL
2	Dimension of Age Group ETL
3	Dimension of Geography ETL
4	Dimension of Treatment Type ETL
5	Dimension of Lab Confirmation ETL

No	ETL Name
6	Dimension of Medication Result ETL
7	Dimension of Spread Starting Point ETL
8	Dimension of Occupation ETL
9	Dimension of Activity Starting Point ETL
10	Dimension of Parasite Type ETL
11	Dimension of Malaria Type ETL
12	Dimension of Medicine Type ETL
13	Dimension of Starting Point Reference ETL


NOVA-EKA.UPT_...ARIA_DW_GALAXY
SQLQuery13.sql - (...va Eka Diana (54))
Results
Messages

```
/****** Script for SelectTopNRows command from SSMS
SELECT TOP 1000 [Hasil Pengobatan]
      ,[HasilKey]
  FROM [UPT_MALARIA_DW].[dbo].[DimHasilPengobatan]
```

	Hasil Pengobatan	HasilKey
1	GPfo	1
2	S	2
3	GPk	3
4	FUTL	4

Fig. 9. Result of DimHasilPengobatan ETL execution

C. *Fact ETL packages*

ETL packages in this group are used to fill in the fact tables by extracting data from dimension and stage tables. Three ETL packages are created to populate data into Fact of Malaria Data, Fact of Area Statistic, and Fact of Logistic. Each of these ETL consists of two components, OLE DB Source and OLE DB Destination. To fill in these fact tables, OLE DB Source is created as JOIN SQL command between dimensional and stage tables. For example, Fig. 10 illustrates SQL command that is used to build OLE DB Source and also the preview result to fill in Fact of Malaria Data table.

D. *Malaria Case Reports*

Malaria data warehouse can be utilized to generate important information for high-level management in the Public Health Department. Users at this level can make a quick and precise decision to overcome Malaria spread by accessing information pooled in the data warehouse.

Fig. 11 depicts a preview of reporting application that use data warehouse information. This figure illustrates the number of Malaria cases based on the type of Malaria Parasite. On January 2013, number of patients in Batulicin1 area diagnosed with Plasmodium Falciparum, Plasmodium Vivax, and Mixed parasite, were 10%, 51%, and 36%, respectively. Based on this information, the health department can decide about what kind of treatments and medicines should be given to those areas. Fig. 12 also gives a preview of reporting application that can be generated from malaria data warehouse by giving some parameters. Here, patients are classified based on their primary occupation. Occupation information is important to find the possible reason of Malaria spread. If most of the patients diagnosed with Malaria have an occupation as farm workers, then the possible reason is the after harvest condition of farming areas that have not been recovered yet. Otherwise, if most of the patients are a miner, then perhaps the responsibility parties have not taken any action to restore the field to its previous condition.

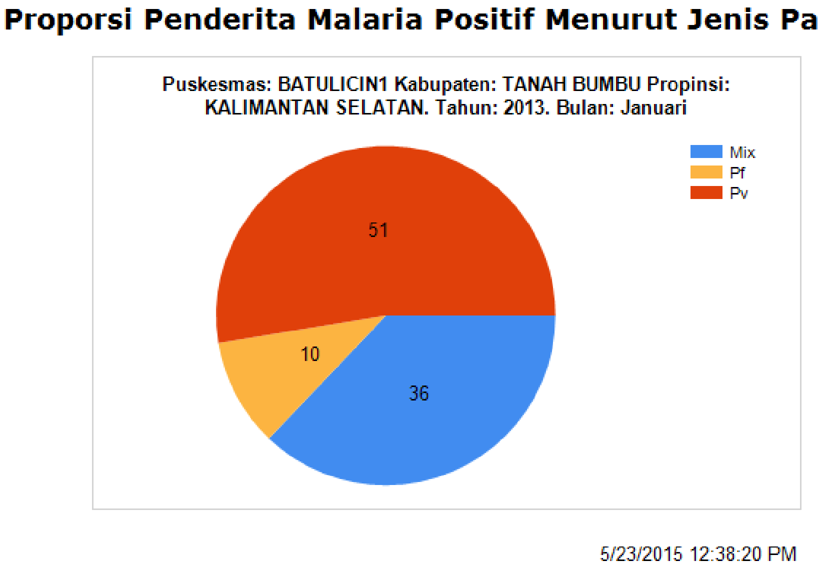
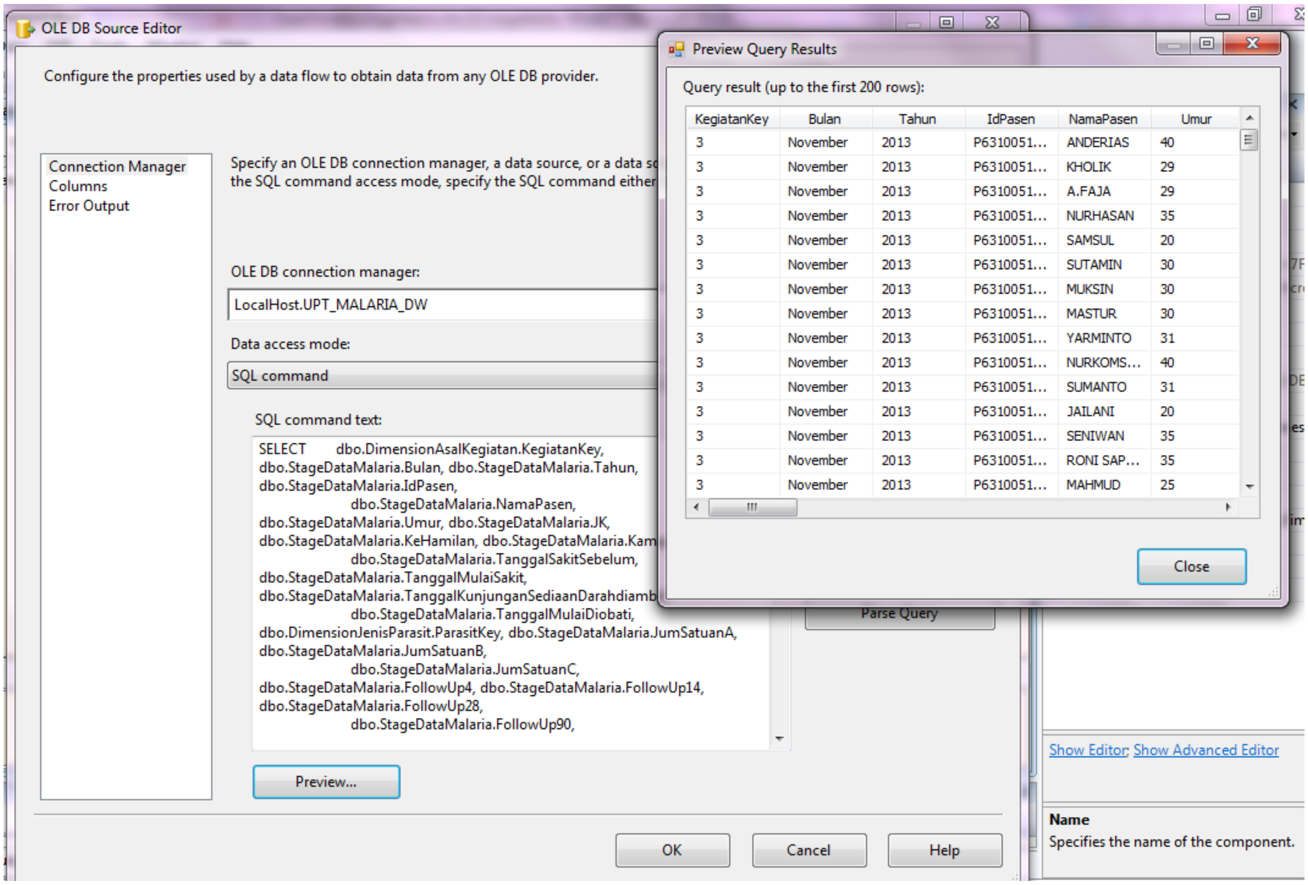


Fig. 10. Result of DimHasilPengobatan ETL execution



Configure the properties used by a data flow to obtain data from any OLE DB provider.

Specify an OLE DB connection manager, a data source, or a data s
the SQL command access mode, specify the SQL command either

OLE DB connection manager: LocalHost.UPT_MALARIA_DW

Data access mode: SQL command

SQL command text:

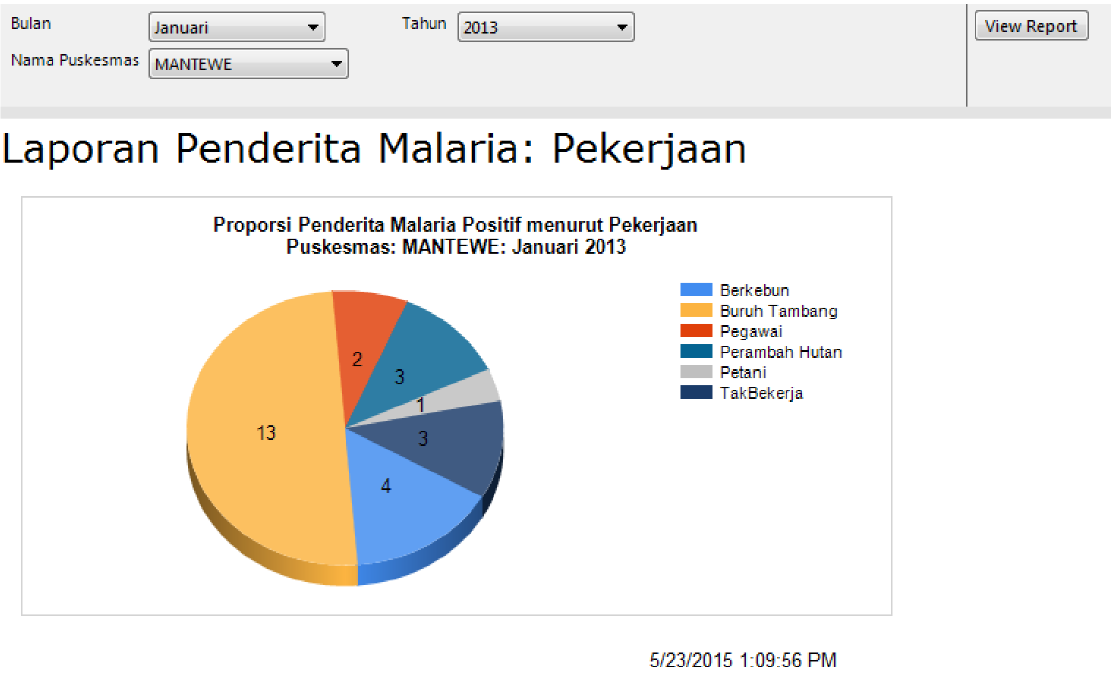
```
SELECT      dbo.DimensionAsalKegiatan.KegiatanKey,
dbo.StageDataMalaria.Bulan, dbo.StageDataMalaria.Tahun,
dbo.StageDataMalaria.IdPasen,
                  dbo.StageDataMalaria.NamaPasen,
dbo.StageDataMalaria.Umur, dbo.StageDataMalaria.JK,
dbo.StageDataMalaria.KeHamilan, dbo.StageDataMalaria.Kam
                  dbo.StageDataMalaria.TanggalSakitSebelum,
dbo.StageDataMalaria.TanggalMulaiSakit,
dbo.StageDataMalaria.TanggalKunjunganSediaanDarahdiamb
                  dbo.StageDataMalaria.TanggalMulaiDiobati,
dbo.DimensionJenisParasit.ParasitKey, dbo.StageDataMalaria.JumSatuanA,
dbo.StageDataMalaria.JumSatuanB,
                  dbo.StageDataMalaria.JumSatuanC,
dbo.StageDataMalaria.FollowUp4, dbo.StageDataMalaria.FollowUp14,
dbo.StageDataMalaria.FollowUp28,
                  dbo.StageDataMalaria.FollowUp90,
```

Query result (up to the first 200 rows):

KegiatanKey	Bulan	Tahun	IdPasen	NamaPasen	Umur
3	November	2013	P6310051...	ANDERIAS	40
3	November	2013	P6310051...	KHOLIK	29
3	November	2013	P6310051...	A.FAJA	29
3	November	2013	P6310051...	NURHASAN	35
3	November	2013	P6310051...	SAMSUL	20
3	November	2013	P6310051...	SUTAMIN	30
3	November	2013	P6310051...	MUKSIN	30
3	November	2013	P6310051...	MASTUR	30
3	November	2013	P6310051...	YARMINTO	31
3	November	2013	P6310051...	NURKOMS...	40
3	November	2013	P6310051...	SUMANTO	31
3	November	2013	P6310051...	JAILANI	20
3	November	2013	P6310051...	SENIWAN	35
3	November	2013	P6310051...	RONI SAP...	35
3	November	2013	P6310051...	MAHMUD	25

Name
Specifies the name of the component.

Fig. 11. OLE DB source & result Preview for Fact DataMalaria



Bulan: Januari
Tahun: 2013
Nama Puskesmas: MANTEWE

Laporan Penderita Malaria: Pekerjaan

5/23/2015 1:09:56 PM

Fig. 12. Reports of Malaria patient’s occupation

VI. Conclusion

In this study, a comprehensive data warehouse is proposed for managing and extracting the salient information about Malaria cases in Tanah Bumbu. This malaria data warehouse is centralized architecture of galaxy or constellation scheme that consists of 3 fact tables and 13 dimension tables. SSIS engine is employed to build ETL packages that load data from various data sources into a stage, dimension, and fact tables in the malaria data warehouse. Timely reports, i.e. number of patients diagnosed with malaria report, can be generated by extracting information from Fact of Malaria Data and Geographic Dimension.

This malaria data warehouse is a foundation to integrate all malaria cases from various endemic and non-endemic areas in Indonesia. Future work would create and generate data mining rules to know the possible starting point where malaria case happened. Hence, the related official government may take an appropriate action to overcome this problem. Furthermore, information existed in the data warehouse is also important for biologists to determine the habitat and the season where mosquito of the genus Anopheles lives. Therefore, the public health could make some preventive actions for minimizing an upcoming number of malaria cases.

References

[1] WHO, “World malaria report 2013,” WHO Global Malaria Programme, World Health Organization, 2013.

[2] Riskesdas, “Riset kesehatan dasar,” Badan Penelitian Dan Pengembangan Kesehatan, Kementrian Kesehatan RI, 2013.

[3] Golfarelli M. & Rizzi S., Data warehouse design: modern principles and methodologies, McGraw-Hill, 2009.

[4] Alazmi A. R., “Data warehousing implementations: a review,” International Journal on Data Mining and Intelligent Information Technology Applications (IJMIA), Vol. 4, No. 1, 2014.

[5] Ramick D.C., “Data warehousing in disease management programs”, J. Healthc Inf Manag, Vol. 15, pp. 15:99-105, 2001.

[6] Inmon W. Building the data warehouse, Wiley Technology Publishing, Wiley, 2005.

[7] Pardillo C.J. and Mazon J.N., “Using ontologies for the design of data warehouses,” Int. J. Database Manag. Syst., vol. 3, no. 2, pp. 73–87, May 2011.

[8] Demarest M., “Understanding the data warehouse lifecycle,” WhereScape White Paper, 2013.

[9] Nilakanta S., Scheibe K. and Rai A., “Dimensional issues in agricultural data warehouse designs,” Computer and Electronics in Agriculture, Vol. 60, No. 2, pp. 263-278, 2008.

[10] Sen A. and Sinha A.P., “A comparison of data warehousing methodologies using a common set of attributes to determine which methodology to use in a particular data warehousing project,” Communications of the ACM, Vol. 48, No. 3, pp. 79-84, 2005.

[11] Singh S. and Malhotra S., “Data warehouse and its methods,” Journal of Global Research in Computer Science, Vol. 2, No. 5, pp.113-115, 2011.

[12] Ariyachandra T. and Watson H.J., “Which data warehouse architecture is most successful?,” Business Intelligence Journal, Vol. 11, No. 1, pp. 4-6, 2006.

[13] Alazmi A. R., “Data warehousing implementations: a review,” International Journal on Data Mining and Intelligent Information Technology Applications (IJMIA), Vol. 4, No. 1, 2014.

[14] Rainardi V., Building a data warehouse with examples in sql server,” Apress, pp. 29-48, 2008.